"Extraction automatique de collocations : Peut-on étendre le test exact de Fisher à des séquences de plus de 2 mots?"

Bestgen, Yves

Abstract

In textual statistics, as in natural language processing and corpus linguistics, the study of sequences of contiguous words that occur together more often than by chance is a major topic of interest. In the case of pairs of words, Fisher's exact test is becoming the reference index to identify them. The objective of this study is to propose a generalization of this index to the analysis of trigrams and longer sequences using a Monte-Carlo procedure. The results of an initial evaluation suggest that this approach could complement other indices, but also that it has a major drawback: a large number of trigrams get a maximum score of collocation.

Document type : *Communication à un colloque (Conference Paper)*

Extraction automatique de collocations : Peut-on étendre le test exact de Fisher à des séquences de plus de 2 mots ?

Yves Bestgen[1]

CECL – Université catholique de Louvain – yves.bestgen@uclouvain.be

Abstract

In textual statistics, as in natural language processing and corpus linguistics, the study of sequences of contiguous words that occur together more often than by chance is a major topic of interest. In the case of pairs of words, Fisher's exact test is becoming the reference index to identify them. The objective of this study is to propose a generalization of this index to the analysis of trigrams and longer sequences using a Monte-Carlo procedure. The results of an initial evaluation suggest that this approach could complement other indices, but also that it has a major drawback: a large number of trigrams get a maximum score of collocation.

Résumé

En statistique textuelle, comme en traitement automatique du langage et en linguistique de corpus, les séquences de mots contigus qui se produisent plus souvent que le hasard ne le prédit deviennent un objet d'étude de plus en plus important. Dans le cas des paires de mots, le test exact de Fisher s'est récemment imposé comme la mesure de référence pour les identifier. L'objectif de cette étude est de proposer une généralisation de ce test à l'analyse des trigrammes et des séquences plus longues au moyen d'une procédure d'estimation de type Monte-Carlo. Les résultats d'une première évaluation indiquent que cette approche est susceptible de complémenter d'autres indices, mais aussi qu'elle présente une limitation importante : l'attribution à un très grand nombre de trigrammes d'un score maximal de collocation.

Mots-clés : linguistique de corpus, n-grammes, collocations, mesures d'association, test de permutation

1. Introduction

Si, en statistique linguistique, le mot est l'unité langagière la plus évidente et la plus fréquemment analysée, les séquences de plusieurs mots sont devenues depuis plus de vingt ans des objets d'études presque aussi importants (Church et Hanks, 1990 ; Daille, 1994 ; Lafon et Salem, 1983 ; Lebart et Salem, 1994). Suivant le point de vue adopté et la discipline scientifique, de nombreuses appellations leur ont été données : segments répétés, n-grammes de mots, collocations, expressions polylexicales (*multiword units*), "paquet de mots" (*lexical bundles*), expressions phraséologiques... Cette diversité d'appellation souligne aussi la nature très différente des expressions en question : idiomes, entités nommées, noms composés, terminologie technique, expressions empruntées à une autre langue... L'intérêt que ce type d'expressions suscite s'explique largement par leur importance dans un grand nombre d'applications en lexicographie bien sûr, mais aussi en statistique textuelle (analyse des réponses à des questions ouvertes, lexicologie politique), en linguistique computationnelle et en recherche d'information (génération automatique de textes, fouille d'opinion), en apprentissage des langues étrangères (évaluation automatique de textes).

[1] Chercheur qualifié du F.R.S-FNRS.

La présente recherche s'intéresse spécifiquement aux séquences de mots contigus (appelées dans la suite *n-grammes*) qui se produisent plus souvent que le hasard ne le prédit (appelées dans la suite *collocations*), une définition très englobante, même si elle laisse de côté les quasi-segments (Becue et Peiro in Lebart et Salem, 1994, p. 124), et largement acceptée[2] en traitement automatique du langage et en linguistique de corpus (Evert, 2009 ; Inkpen et Hirst, 2002 ; McKeown et Radev, 1999). Cette définition impose l'étude de grands corpus de textes afin de distinguer non seulement ce qui est une expression collocationnelle de ce qui est juste une combinaison acceptable de mots, mais aussi les expressions fortement collocationnelles de celles qui ne le sont que faiblement. Pour prendre de telles décisions, la fréquence du n-gramme est évidemment importante, mais elle ne peut suffire : des mots très fréquents ont une probabilité non négligeable de se suivre dans un corpus. Le tableau 1 illustre une telle situation. Les dénombrements pour deux bigrammes dans le corpus TASA (Touchstone Applied Science Associates, Inc.), un corpus de référence en anglais américain qui compte presque 15 millions d'occurrences, sont présentés sous la forme classique de tables de contingence 2×2. Les lignes de ces tables représentent la présence ou l'absence du premier mot du bigramme et les colonnes celles du second mot. La cellule *a* indique donc le nombre de fois que le bigramme a été observé dans le corpus. La cellule *b* indique le nombre de fois qu'un bigramme commençant par le premier mot, mais ne se terminant pas par le second a été observé et ainsi de suite.

	Mot 2		
Mot 1	the	¬the	Total
you	***a*** 351	***b*** 104 132	104 483
¬you	***c*** 873 150	***d*** 13 915 461	14 788 611
Total	873 501	14 019 593	14 893 094

	Mot 2		
Mot 1	than	¬than	Total
larger	350	1 674	2 024
¬larger	20 561	14 870 509	14 891 070
Total	20 911	14 872 183	14 893 094

Tableau 1. Dénombrements pour deux bigrammes dans le corpus TASA

Dans ce corpus, le bigramme *you_the*, que l'on trouve par exemple dans *Suppose your teacher gives you the task...* (TASA, ID=Agatha18.01.02) et qu'il est difficile de considérer comme collocationnel, est observé 351 fois, soit une occurrence de plus que le bigramme *larger than* qui correspond aux définitions classiques des expressions collocationnelles (McKeown et Radev, 1999). Comparé aux fréquences de *more* et de *than*, 350 est un nombre élevé. Par contre, comparé à celles de *you* et de *the*, 351 est très faible. Il est donc nécessaire de prendre en compte la probabilité que les mots qui composent un n-gramme ont de se suivre dans le corpus afin de pouvoir déterminer si la fréquence d'occurrence du n-gramme est supérieure à ce à quoi on s'attendrait si les mots étaient disposés aléatoirement.

Dans le cas particulier des bigrammes, cette question a retenu l'attention de nombreux chercheurs et un grand nombre de mesures ont été proposées, le plus souvent empruntées à d'autres disciplines dans lesquelles les coefficients d'association et les tests inférentiels pour les tables de contingence 2×2 jouent un rôle important. Le grand nombre d'indices proposés, (Pecina, 2010) en a rassemblé plus de 50, trouve largement son origine dans la diversité de leurs champs d'applications, mais aussi dans l'étendue des cas à prendre en compte : des mots et des expressions extrêmement rares et d'autres très fréquents. Ces dernières années, un

[2] Même si c'est loin d'être la seule puisque, plutôt que de s'intéresser aux contigüités immédiates, on peut considérer les cooccurrences dans une fenêtre de mots ou dans une unité plus grande comme la phrase ou le paragraphe.

indice s'est imposé comme l'indice de référence, utilisé par exemple comme critère pour évaluer les autres indices (Evert, 2009 ; Gries et Stefanowitsch, 2004 ; Moore, 2004) : le test exact de Fisher, basé sur la distribution hypergéométrique chère à nombre de chercheurs en statistique textuelle (p.e., Brunet, 2007 ; Lafon, 1981 ; Heiden, 2004).

Les mesures d'association pour les n-grammes de plus de 2 mots ont beaucoup moins retenu l'attention, les travaux se limitant le plus souvent à proposer différentes manières de généraliser les mesures conçues pour les bigrammes. Ce manque d'études conduit Evert (2009), dans son chapitre de synthèse sur les collocations, à les placer sur la liste des choses à faire, écrivant : "*It is therefore desirable to develop suitable measures for word triples and larger n-tuples*." Un constat identique est fait par Nerima et al. (2010). Ce manque d'études est d'autant plus problématique qu'on peut penser que la prise en compte des expressions de plus de 2 mots est susceptible de grandement faciliter le traitement des bigrammes (Lyse et Andersen, 2012). Nombre de bigrammes sont des tronçons d'expressions plus longues. Les analyser à ce niveau plus global devrait permettre de les identifier plus facilement, mais aussi d'en expurger la liste des bigrammes de sorte que celle-ci ne contient plus que de vrais bigrammes.

Face à ce constat, l'objectif de la présente étude est de proposer une procédure pour généraliser le test de Fisher **lorsqu'il est employé comme mesure d'association collocationnelle** à des n-grammes de plus de deux mots au moyen d'une procédure d'estimation de type Monte-Carlo. La suite du texte est structurée de la manière suivante. Après avoir brièvement décrit le test exact de Fisher, la section 2 explique à quoi il correspond dans le cas de l'étude des séquences de deux mots contigus. Elle se conclut en déduisant une approche pour le généraliser aux n-grammes de plus de 2 mots. La troisième section propose une implémentation de cette approche et la quatrième section applique cette procédure aux bigrammes et aux trigrammes afin de proposer une première évaluation de son intérêt. La conclusion souligne la limite principale de l'approche proposée.

2. Le test de Fisher et son emploi comme mesure d'association de mots

Les principales mesures d'association employées dans l'étude des collocations, et plus particulièrement des bigrammes, ont pour objectif de déterminer si la fréquence d'occurrence du bigramme est supérieure à ce à quoi on s'attendrait **si les mots étaient disposés aléatoirement dans le corpus**. Cette condition correspond à l'hypothèse d'indépendance entre les deux mots qui composent le bigramme ; la fréquence attendue (A) est alors égale à (voir tableau 1) :

$$A = \frac{(a+b)(a+c)}{(a+b+c+d)} = \frac{Freq_{Mot1} \times Freq_{Mot2}}{Freq_{Tot}}$$

Il est important de souligner, même si ceci est connu de nombre de praticiens de la statistique textuelle, que prendre comme point de repère l'hypothèse simpliste que les mots d'un corpus puissent être disposés aléatoirement est un raccourci évidemment fallacieux qui peut mener à une sacralisation de la modélisation probabiliste en statistique textuelle (voir pour une critique récente, (Bestgen, 2014)). L'hypothèse d'indépendance doit donc être vue comme une hypothèse de commodité dont le seul but est de permettre l'attribution aux bigrammes de scores dont seul l'ordre de grandeur traduit l'intérêt pour des analyses plus approfondies. C'est, entre autres, pour cette raison que les probabilités obtenues dans la suite par l'entremise du test exact de Fisher sont appelées **scores de collocation**.

La figure 1 présente quelques indices parmi ceux qui sont les plus souvent employés (Evert, 2009 ; Granger et Bestgen, sous presse ; Khokhlova, 2008 ; Manning et Schütze, 1999). Les indices *pMI* (pour *pointwise Mutual Information*) et *t*, popularisés dans ce champ de recherche par Church et Hanks (1990), comparent la fréquence observée (*O*) du bigramme (cellule *a* dans le tableau 1) à la fréquence attendue. Ils sont fréquemment employés en lexicographie et en linguistique de corpus malgré leurs lacunes : *pMI* est connu pour attribuer des scores très élevés à des n-grammes composés de mots très rares et Evert (2004) a montré que la dérivation de l'indice *t* du test *t* de Student était incorrecte.

$$pMI = \log_2\left(\frac{O}{A}\right) \qquad t = \frac{O-A}{\sqrt{O}} \qquad G^2 = \sum_{ij} O_{ij} \log\left(\frac{O_{ij}}{A_{ij}}\right)$$

Figure 1. Formules de trois indices d'association fréquemment employés

(Dunning, 1993) a recommandé l'emploi de G^2, le Khi2 du maximum de vraisemblance, qui est un test inférentiel évaluant le degré d'accord entre les fréquences observées et les fréquences attendues des quatre cellules de la table sous l'hypothèse d'indépendance. G^2 est toutefois inadéquat lorsqu'au moins une des fréquences attendues dans le tableau est très inférieure à 1, une situation fréquente dans ce genre d'études en raison de la rareté d'un grand nombre des mots qui composent un corpus (Pedersen, 1996). Pedersen et al. (1996) ont proposé[3] de remédier à ce problème en employant le test exact de Fisher. La proposition de Fisher est de considérer l'ensemble des tables de contingence qu'il est possible de construire en respectant les totaux marginaux réellement obtenus et de déterminer la proportion de celles-ci qui donne lieu à un résultat au moins aussi extrême que celui observé (Agresti, 2007, pp. 45-47), une table plus extrême étant une table dans laquelle la fréquence du bigramme est plus élevée que celle observée (cellule *a*). La formule de calcul est la suivante (Evert, 2004) :

$$PFisher = \sum_{i=a}^{\min(a+b,a+c)} \frac{\binom{(a+c)}{i}\binom{(b+d)}{(a+b-i)}}{\binom{(a+b+c+d)}{(a+b)}}$$

Comme indiqué dans l'introduction, ce test est de plus en plus populaire dans les travaux sur les collocations. À ma connaissance, il n'a jusqu'à présent été appliqué qu'à l'étude des bigrammes, dont les données peuvent être représentées par des tables de contingence à deux dimensions. Or, l'analyse de collocation de 3 mots ou plus impose la construction de tables à 3 dimensions ou plus. G^2 se généralise très facilement à ce genre de tables puisqu'il est à la base des modèles log-linéaires. Toutefois, on peut se demander pourquoi il serait plus adéquat pour l'étude des trigrammes et des séquences plus longues que pour les bigrammes.

L'objectif de la présente étude est de proposer une généralisation du test exact de Fisher à l'analyse des trigrammes et des séquences plus longues au moyen d'une procédure d'estimation de type Monte-Carlo. Il est important de souligner que ce travail n'essaye pas de généraliser le test exact de Fisher à des tables à plus de deux dimensions, mais seulement à l'analyse des trigrammes, quadrigrammes... qui peuvent être représentés par ce genre de tables. On notera aussi que des tests exacts (ou des approximations) ont été proposés pour des tables de contingence à trois ou plus dimensions (Zeltermanab et al., 1995). Les procédures

[3] Mais voir Lafon (1981) pour une application de la même loi hypergéométrique au nombre de rencontres de deux mots dans un espace limité, le plus souvent la phrase.

proposées ne sont toutefois pas adaptées à l'étude des n-grammes en raison de la taille de l'échantillon et du grand nombre de tests à effectuer et parce qu'on souhaite tester une hypothèse très spécifique et directionnelle (la probabilité d'avoir au moins autant d'occurrences de ce trigramme par le seul fait du hasard).

Le point de départ de cette généralisation se situe dans une méthode alternative de calcul de la probabilité issue du test exact de Fisher dans le cas des bigrammes. Classiquement, cette probabilité est obtenue au moyen de la formule donnée ci-dessus, mais une autre approche est possible : employer une procédure de permutation de type Monte-Carlo et donc générer un échantillon aléatoire des tables de contingence possibles (étant donné les totaux marginaux) et déterminer la proportion de celles-ci qui donnent lieu à une valeur au moins aussi extrême que celle observée. Il s'agit de la procédure proposée par (Pedersen et al., 1996) dans leur article recommandant l'emploi du test de Fisher pour l'étude des bigrammes ; elle est aussi employée par les logiciels de statistique, comme SAS, pour les situations dans lesquelles l'approche exacte est computationnellement trop longue à mettre en œuvre. Pour générer ces tables, on peut employer, dans le cas des collocations, les procédures habituelles pour n'importe quelle table de contingence. Mais on peut aussi s'intéresser à ce à quoi correspond vraiment cette procédure de permutation dans le cas des n-grammes. Un corpus étant une longue séquence de formes graphiques, une permutation consiste à mélanger ces formes de manière aléatoire et à compter le nombre de fois qu'un bigramme donné est observé dans cette permutation. Chaque permutation génère donc une table de contingence aléatoire. Cette procédure d'estimation se généralise aisément aux séquences de plus de deux mots puisqu'il suffit de compter dans chaque permutation aléatoire, non seulement les bigrammes, mais aussi les trigrammes, quadrigrammes... Les valeurs obtenues sont comparées aux fréquences observées dans le corpus original et permettent donc de déterminer la probabilité d'obtenir une fréquence au moins aussi extrême que celle observée par le seul effet du hasard.

3. Implémentation

L'objectif est de déterminer la probabilité qu'a le hasard de produire au moins autant d'occurrences d'un n-gramme que le nombre réellement observé dans le corpus. Pour ce faire, une procédure classique de test de permutation de type Monte-Carlo a été employée. Elle est composée de deux étapes qui sont répétées un grand nombre de fois :

- Permutation des occurrences (*tokens*) qui composent le corpus
- Pour chaque n-gramme présent dans le corpus original, déterminer si sa fréquence dans la permutation est au moins égale à sa fréquence dans le corpus original. Si c'est le cas, on incrémente de 1 le compteur correspondant à ce n-gramme.

Lorsque le nombre d'itérations souhaité est atteint, on divise, pour chaque n-gramme, le nombre contenu dans le compteur par le nombre d'itérations effectuées ; on obtient ainsi une estimation de la probabilité qu'a le hasard de produire au moins autant d'occurrences de ce n-grammes que le nombre réellement observé.

Concrètement, les permutations ont été effectuées au moyen de l'algorithme de Fisher-Yates tel qu'implémenté par Durstenfeld (Knuth, 1998). Il garantit que toutes les permutations ont la même probabilité d'être produites et nécessite un temps proportionnel au nombre d'éléments à permuter. Pour générer les nombres (pseudo-)aléatoires nécessaires, j'ai employé l'algorithme JKISS de Jones (2010) dont la période est approximativement égale à 2^{127} et qui passe avec succès les tests les plus complexes visant à démontrer que la séquence générée est une "*good imitation of a sequence of independent uniform random variables*"

(L'ecuyer et Simard, 2007). L'étape de comptage des fréquences des n-grammes dans la permutation a été optimisée par le recours à une table de hachage, telle qu'implémentée dans un ensemble de macros pour le langage *C* par Hanson (2013) ; elle permet un accès direct aux n-grammes présents dans le corpus original.

Étant donné qu'on s'intéresse à des n-grammes de différentes longueurs, deux approches sont possibles : extraire des séquences d'une seule longueur à la fois et donc relancer l'ensemble de la procédure pour chaque nouvelle longueur ou extraire toutes les longueurs souhaitées en une seule fois. La deuxième approche a été choisie parce qu'elle est plus économique en temps. Elle accroît cependant l'espace mémoire nécessaire puisqu'il faut stocker tous les n-grammes de chaque longueur présents dans le corpus original. De plus, elle rend problématique l'emploi d'un tableau de hachage pour chaque longueur en raison de l'espace mémoire qu'un tel tableau requiert. La solution adoptée consiste à n'employer la table de hachage que pour les séquences de deux mots et à utiliser une recherche binaire pour les tailles plus longues. Afin d'accélérer la recherche, on a stocké pour chaque n-gramme d'une longueur *n* un pointeur vers le premier n-gramme d'une longueur *n+1* qui commence par cette séquence. On notera que si un n-gramme d'une longueur *n* présent dans la permutation n'existe pas dans le corpus original, il est inutile de chercher les n-grammes de longueur *n+1*, *n+2*...

4. Evaluation de la procédure proposée

Afin de pouvoir se faire une première idée de l'intérêt de la procédure proposée, celle-ci a été appliquée à un corpus de presque 15 millions d'occurrences et les deux analyses suivantes ont été effectuées :

- comparer pour les bigrammes les probabilités obtenues au moyen de la procédure de permutation aux probabilités exactes fournies par le test de Fisher.
- effectuer une première évaluation de l'efficacité de la procédure dans le cas des n-grammes de plus de deux mots en confrontant les trigrammes sélectionnés par celle-ci à ceux sélectionnés par deux autres indices d'association : *pMI* et *t*.

4.1. Méthode

4.1.1. Corpus

Le corpus *TASA - General Reading up to 1st year college* est composé d'extraits de textes, d'une longueur moyenne approximative de 250 mots, obtenus par un échantillonnage aléatoire des textes que lisent les élèves et les étudiants américains. La version à laquelle T.K. Landauer (Institute of Cognitive Science, University of Colorado, Boulder) m'a donné accès contient 44 486 documents. Ce corpus a été segmenté au moyen du logiciel *TreeTagger* (Schmid, 1994). Toutes les formes graphiques (mots, mais aussi nombres, symboles et signes de ponctuation) détectées par le logiciel ont été conservées, soit un peu moins de 15 millions d'occurrences (*tokens*) dont approximativement 12 millions de mots.

4.1.2. Obtention des probabilités

La procédure d'estimation décrite à la section 3 a été programmée en *C* sur un iMac Intel Core i5 2.66 Ghz. Le programme emploie 345Mb de mémoire réelle pour traiter le corpus TASA qui contient 2 210 185 séquences différentes de deux mots et 7 229 912 séquences différentes de trois mots. Une itération prend 5,54 secondes. Si on initialise différemment le

générateur pseudo-aléatoire, plusieurs exemplaires du programme peuvent fonctionner simultanément. Au total, 400 000 itérations ont été effectuées.

4.2. Analyses et résultats

4.2.1. Comparaison avec le test exact de Fisher dans le cas des bigrammes

La probabilité exacte (issue du test de Fisher, abrégée en *PFisher*) d'obtenir par le seul fait du hasard autant d'occurrences de chaque bigramme a été comparée à l'estimation dérivée des permutations (*PPerm* dans la suite). Comme attendu, ces deux ensembles de valeurs sont presque identiques. Une régression linéaire (avec *PFisher* comme variable dépendante) produit une ordonnée à l'origine de 5.3^{-7} et une pente égale à 1. Le R^2 est égal à 0,999998.

Une mesure plus fine du degré d'accord entre les deux mesures consiste à comparer les décisions prises par celles-ci pour chaque bigramme lorsqu'on emploie pour chacune des deux mesures le même seuil de probabilité (α) pour décider si un bigramme est une collocation ou non. Le tableau 2 donne le pourcentage de décisions identiques pour différents seuils. Comme on peut le voir, l'accord est très bon. On note toutefois une diminution importante de celui-ci à partir d'un seuil de 0,0000001, c'est-à-dire lorsque *PFisher* doit être inférieur à 1/400 000 (0,0000025). Dans ces cas-là, *PPerm* est égale[4] (au mieux) à 0. Au total, on a observé 182 499 bigrammes avec *PPerm*=0, soit 8,26 % des bigrammes. Par contre, la procédure *Freq* de *SAS*, employée pour calculer *PFisher*, n'a rapporté que 2 294 valeurs nulles (c'est-à-dire arrondies à 0 parce que trop petites pour être représentées avec une précision suffisante, le seuil se situant vers 1^{-304}). Cette discordance trouve donc son origine dans le nombre de permutations effectuées, trop petit pour estimer de telles probabilités. Cette limitation de l'approche deviendra encore plus manifeste lorsque les trigrammes seront considérés.

Seuil de décision					
0,05	0,01	0,001	0,00001	0,000001	0,00000001
99,88 %	99,86 %	99,78 %	99,45 %	98,18 %	96,54 %

Tableau 2. Pourcentage de décisions identiques par PFisher et PPerm selon le seuil α employé

4.2.2. Application aux trigrammes

Pour évaluer l'efficacité d'un indice d'association, la solution la plus évidente consiste à confronter les décisions basées sur cet indice avec une liste de vraies collocations. Hélas, une telle liste n'existe pas. Le caractère fortement lacunaire des expressions reprises dans *WordNet* a été fréquemment souligné y compris par les chercheurs qui ont utilisé cette ressource (Petrovic et al., 2010 ; Schone et Jurafsky, 2001). Une solution plus rigoureuse consiste à demander à des lexicographes de constituer cette liste sur la base d'un échantillon aléatoire des n-grammes présents dans le corpus, comme l'a fait (Pecina, 2010) pour le tchèque. Constituer une telle norme est cependant inapplicable dans la présente étude exploratoire d'autant plus qu'un nombre important d'experts est nécessaire pour que la

[4] En toute rigueur, on ne devrait pas attribuer une valeur minimale à *PPerm* de 0 puisque, comme l'ont souligné entre autres (Knijnenburg et al., 2009), les probabilités issues d'un test de permutation (et d'un test exact) ne peuvent jamais être égales à 0 (au minimum le score réellement observé est possible), mais, dans le cas présent, il me semble que la simplification qui consiste à estimer *PPerm* au moyen du nombre de permutations ayant donné une valeur au moins aussi grande divisé par le nombre total de permutations facilite l'exposé.

fiabilité soit acceptable. Aussi ai-je opté pour une approche similaire à celle employée pour l'analyse des bigrammes : situer le nouvel indice par rapport aux autres indices employés pour l'étude des séquences de plus de deux mots. Il s'agit donc de déterminer si les trigrammes considérés comme fortement collocationnels par *PPerm* le sont aussi par au moins un autre indice.

Pour ces analyses, les indices d'association *pMI* et *t* ont été sélectionnés[5] en raison de leur popularité dans les études des bigrammes ainsi que dans celles des séquences plus longues (Durrant et Schmitt, 2009 ; Ellis et al., 2008 : Evert, 2010). Les formules de ces indices pour les trigrammes sont identiques à celles employées pour les bigrammes (voir figure 1), la fréquence attendue étant calculée au moyen d'une simple extension de la formule classique comme implémentée dans le *Ngram Statistics Package*[6] (Banerjee et Pedersen, 2003) :

$$A_{Mot1,Mot2,Mot3} = \frac{Freq_{Mot1} \times Freq_{Mot2} \times Freq_{Mot3}}{Freq_{Tot} \times Freq_{Tot}}$$

Étant donné que le test de permutation attribue un score de collocation maximal identique (une probabilité de 0, mais voir note 2) à 1 142 188 trigrammes sur les 7 229 912, on a dichotomisé l'ensemble des trigrammes selon que *PPerm* est égal ou non à 0. On obtient donc une catégorie contenant les bons trigrammes et une catégorie contenant les moins bons ou les mauvais trigrammes. Pour les autres indices, on a considéré comme étant de bons trigrammes les 1 142 188 trigrammes ayant les scores les plus élevés. De cette manière, chaque indice classe un même nombre de trigrammes dans les deux catégories.

		PPerm				Total	
		Non		Oui			
pMI	t	N	%	N	%	N	%
Non	Non	5042044	69,74	1	0,00	5042045	69,74
	Oui	422810	5,85	622869	8,62	1045679	14,46
Oui	Non	622863	8,62	422816	5,85	1045679	14,46
	Oui	7	0,00	96502	1,33	96509	1,33
Total		6087724	84,20	1142188	15,80	7229912	100

Tableau 3. Répartition croisée des trigrammes selon qu'ils ont été sélectionnés par chacun des trois indices : pMI, t et PPerm (fréquence et pourcentage du total)

Le tableau 3 donne la répartition croisée des trigrammes dans ces deux catégories pour les trois indices. On observe que 5 042 044, soit 69.74% du total des trigrammes, ne sont sélectionnés par aucun indice et 96 502, soit 1.33%, sont sélectionnés par les trois indices. On observe aussi qu'un seul trigramme a été sélectionné par le seul indice *PPerm*. Il s'agit de *that sells bags* qu'il est difficile de considérer comme un bon trigramme. Par contre, 622 863

[5] Initialement, il était prévu d'employer aussi G^2 calculé sur la base du modèle d'indépendance complète. Cet indice présente toutefois un défaut qui le rend inadéquat pour l'étude : il teste une hypothèse non-directionnelle et très générale. Il s'ensuit qu'il attribue une valeur élevée que la fréquence observée du n-gramme soit beaucoup plus grande ou beaucoup plus petite que la fréquence attendue. Cet indice donne également une valeur très élevée à un trigramme comme *family of the*, médiocre pour d'autres indices, parce qu'il inclut une paire de mots (*of the*) qui est très collocationnelle plus un mot quelconque.

[6] http://search.cpan.org/dist/Text-NSP/lib/Text/NSP/Measures/3D/MI/pmi.pm

trigrammes sélectionnés par *pMI* ne le sont par aucun des deux indices. Pour *t*, le nombre correspondant est 422 810. Seuls 7 trigrammes sont sélectionnés conjointement par *pMI* et *t*, mais non par *PPerm*. Il s'agit de *at least nominal, are short-day plants, best society unless, compensate for wear, emma had agreed, north american review, nuisances that should*.

Il apparaît donc que (presque) tous les trigrammes sélectionnés par l'indice *PPerm* sont au moins sélectionnés par un des deux autres indices alors que c'est loin d'être le cas pour les deux autres indices. Ceci suggère que *PPerm*, malgré le fait qu'il attribue un score maximum à de nombreux n-grammes, pourrait être utile pour filtrer les collocations sélectionnées par un des deux autres indices. Il est toutefois nécessaire d'analyser plus finement ces résultats afin de déterminer si les trigrammes qui sont sélectionnés par *PPerm* et un autre indice ne sont pas systématiquement ceux qui ont obtenu le plus haut score sur cet autre indice. Dans un tel cas, PPerm serait évidemment de peu d'utilité puisqu'il suffirait d'éliminer les trigrammes ayant les scores les plus bas pour obtenir le même filtrage.

La figure 2 présente la dispersion des scores les plus élevés pour *pMI* sous la forme d'un histogramme dans lequel on a distingué les trigrammes qui ont été sélectionnés par *PPerm* (en vert) de ceux qui ne l'ont pas été (en rouge : la position dans une barre est sans signification). Comme on peut le voir si *PPerm* approuve majoritairement les trigrammes ayant reçu les scores *pMI* les plus élevés, il en rejette un nombre non négligeable pour privilégier des trigrammes ayant des scores plus faibles. La figure 2 donne la même information pour le test *t*. Si l'effet est nettement moins fort, il est néanmoins visible dans la zone [2.5, 4.0][7].

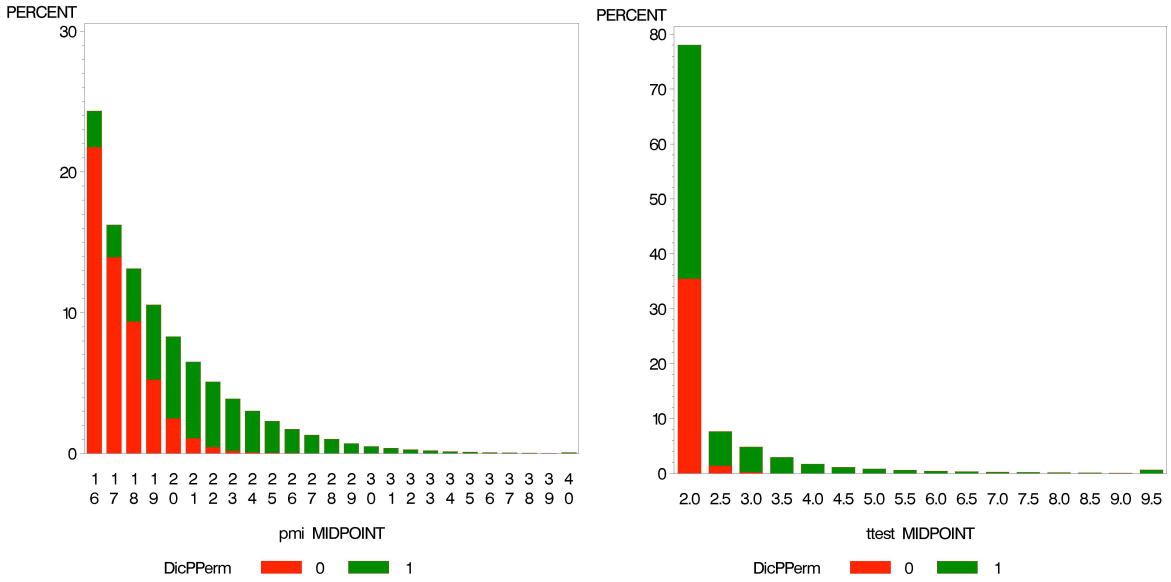

Figure 2. Dispersion des scores les plus élevés pour pMI (à gauche) et t (à droite)

Plus qualitativement, le tableau 4 donne des exemples des trigrammes qui ont obtenu les scores les plus élevés pour un des deux autres indices et qui n'ont pas été sélectionnés par *PPerm*. D'une manière générale, peu d'entre eux peuvent être qualifiés de bonnes collocations.

[7] Ces scores *t* peuvent sembler très faibles, mais ils dépassent le seuil habituellement employé (2) pour considérer qu'un bigramme est une collocation (Durrant et Schmitt, 2009).

Trigramme	*pMI*
genuine korean apothecary	30.81
sexism emphasizes sexual	30.42
shell electron-pair repulsion	30.06
sledgehammer waves crashed	29.31
sexagesimal system survives	29.04
convicted burglar serving	28.65
american college-university theatre	28.38
imposing official residences	28.09
technological advances revolutionized	28.06
hundreds milled aimlessly	28.06

Trigramme	*t*
to it that	4.04
the good of	3.84
we had the	3.82
not on the	3.72
water and the	3.72
and the more	3.69
and only a	3.69
this was to	3.68
been for the	3.66
to the most	3.66

Tableau 4. Trigrammes ayant obtenu les scores les plus élevés selon pMI ou t, mais n'ayant pas été sélectionnés par PPerm

5. Conclusion

À la question posée dans le titre *Peut-on étendre le test exact de Fisher à des séquences de plus de deux mots ?*, la réponse me semble être *Oui, mais*. Oui, parce que la procédure proposée, qui est une approximation de type Monte-Carlo du test exact de Fisher, est applicable aux n-grammes de plus de deux mots et qu'elle n'attribue des scores d'association élevés qu'aux trigrammes qui reçoivent des scores élevés pour d'autres indices alors que ces autres indices ont tendance à sélectionner une large proportion de trigrammes qu'aucun des autres indices analysés ne sélectionne.

Mais, il y a un mais. Cet indice attribue à un très grand nombre de trigrammes un score maximal de collocation (c'est-à-dire la plus petite probabilité pouvant être rapportée étant donné le nombre de permutations effectuées) ; ceux-ci ne peuvent donc être distingués les uns des autres. Ce résultat n'est en soi pas étonnant étant donné qu'un grand nombre de bigrammes sont déjà extrêmement peu probables pour le test exact de Fisher. Il n'empêche que l'indice *PPerm* se révèle trop peu discriminatif. Est-il imaginable de dépasser cette limitation? Il est évidemment toujours possible d'effectuer un plus grand nombre de permutations, mais cela semble une approche illusoire étant donné le temps qui serait nécessaire pour effectuer ne fut-ce qu'un milliard de permutations, soit plus de 175 ans sur un noyau du processeur employé dans les simulations rapportées ci-dessus, alors que cela ne correspondrait qu'à une probabilité minimale de 0.000000001, bien loin du 1^{-303} qui peut être obtenu sans difficulté avec le test exact de Fisher lorsqu'il est appliqué aux bigrammes. Une autre approche consisterait à essayer d'employer les travaux de (Knijnenburg et al., 2009). Ces auteurs ont proposé de réduire le nombre de permutations nécessaire pour estimer une probabilité issue d'un test de type Monte-Carlo en approximant la distribution des valeurs extrêmes obtenues lors des permutations au moyen d'une distribution de Pareto généralisée. Il n'est cependant pas évident que cette approche puisse être efficace dans le cas présent en raison de la nature discrète des valeurs de test issues des permutations.

Références

Agresti A. (2007). An introduction to categorical data analysis. Wiley.

Banerjee S. and Pedersen T. (2003). The design, implementation, and use of the Ngram statistics package. *Proceeding of CICLing'03 (4th international conference on Computational linguistics and intelligent text processing)*, pp. 370-381.

Bestgen Y. (2014). Inadequacy of the chi-squared test to examine vocabulary differences between corpora. *Literary and Linguistic Computing*, Advance Access. doi: 10.1093/llc/fqt020.

Brunet E. (2007). Fréquences et séquences. Mise en œuvre dans Hyperbase. *Lexicometrica : Topographie et topologie textuelles*, 7, 20 p.

Church K. and Hanks P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics*, 16: 22–29.

Daille B. (1994). Study and implementation of combined techniques for automatic extraction of terminology. *Proceeding of The Balancing Act Workshop: Combining Symbolic and Statistical Approaches to Language*, pp. 29–36.

Dunning T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19: 61–74.

Durrant P. and Schmitt N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL,* 47: 157-177.

Ellis N., Simpson-Vlach R. and Maynard C. (2008). Formulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpus Linguistics, and TESOL. *TESOL Quarterly,* 42: 375-396.

Evert S. (2004). The Statistics of Word Cooccurrences: Word Pairs and Collocations. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

Evert S. (2009). Corpora and collocations. In Anke Ludeling and Merja Kytö (eds.), *Corpus Linguistics. An International Handbook* (pp. 1211-1248). Mouton de Gruyter.

Evert S. (2010). Google Web 1T 5-Grams Made Easy (but not for the computer). *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*, pp. 32–40.

Granger S. & Bestgen Y. (in press). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *IRAL: International Review of Applied Linguistics in Language Teaching*.

Gries S. and Stefanowitsch A. (2004). Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics*, 9: 97-129.

Hanson T. (2013). uthash: a hash table for c structures. http://troydhanson.github.io/uthash/index.html, version 1.9.8.

Heiden S. (2004). Interface hypertextuelle à un espace de cooccurrences : implémentation dans Weblex. *JADT 2004 : 7es Journées internationales d'Analyse statistique des Données Textuelles,* pp. 577-588.

Inkpen D. and Hirst G. (2002). Acquiring collocations for lexical choice between near synonyms. *Proceedings of SIGLEX workshop on unsupervised lexical acquisition*, pp. 67-76.

Jones D. (2010). Good Practice in (Pseudo) Random Number Generation for Bioinformatics Applications. Departments of Computer Science & Structural and Molecular Biology, University College London. (http://www.cs.ucl.ac.uk/staff/d.jones/GoodPracticeRNG.pdf).

Khokhlova M. (2008). Extracting collocations in Russian: Statistics vs. Dictionary. *JADT 2008 : 9es Journées internationales d'Analyse statistique des Données*, pp. 613-624.

Knijnenburg T., Wessels L., Reinders M. and Shmulevich I. (2009). Fewer permutations, more accurate P-values. *Bioinformatics*, 25: 161–168.

Knuth D. (1998). *The Art of Computer Programming* (vol.2). Addison Wesley Longman.

L'ecuyer P. and Simard R. (2007). TestU01: A C library for empirical testing of random number generators. *ACMTrans. Math. Softw.*, 33, 40 pages.

Lafon P. (1981). Analyse lexicométrique et recherche des cooccurrences. *Mots*, 3: 95-148.

Lafon P. et Salem A. (1983). L'inventaire des segments répétés d'un texte. *Mots*, 6: 161-177.

Lebart L. et Salem A. (1994). *Statistique textuelle*. Dunod.

Lyse G. and Andersen G. (2012). Collocations and statistical analysis of n-grams: Multiword expressions in newspaper text. In G. Andersen (ed.), *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian*, (pp. 79–110).

Manning C. and Schütze H. (1999). *Foundations of Statistical Natural Language Processing*. MIT.

McKeown K. and Radev D. (1999). Collocations. In R. Dale, H. Moisl, and H. Somers (eds.), *A Handbook of Natural Language Processing*. Marcel Dekker.

Moore R. (2004). On log-likelihood-ratios and the significance of rare events. *Proceedings of EMNLP 2004*, pp. 333-340.

Nerima L., Wehrli E. and Seretan V. (2010). A Recursive Treatment of Collocations. *LREC 2010*, pp. 634-638.

Pecina P. (2010). Lexical association measures and collocation extraction. *Language Resources & Evaluation*, 44: 137–158.

Pedersen T. (1996). Fishing for exactness. *Proceedings of the South Central SAS Users Group*. pp. 188-200.

Pedersen T., Kayaalp M. and Bruce R. (1996). Significant lexical relationships. *Proceedings of the 13th National Conference on Artificial Intelligence*, pp. 455-460.

Petrovic S., Snajder J. and Basic B. (2010). Extending lexical association measures for collocation extraction. *Computer Speech and Language*, 24: 383-394.

Schmid H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees, *Proceedings of the International Conference on New Methods in Language Processing*, pp. 44-49.

Schone P. and Jurafsky D. (2001). Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords a Solved Problem? *Proceedings of Empirical Methods in Natural Language Processing*, pp. 100-108.

Zeltermanab D., Chanac I. and Mielke P. (1995). Exact Tests of Significance in Higher Dimensional Tables. *The American Statistician*, 49: 357-361.